

Patrick Niehaus

HOMETOWN

Hilton Head Island, South Carolina

MAJORS

Computational Science, at University of South Carolina Beaufort

ACADEMIC LEVEL

Senior

RESEARCH MENTOR

Mark Holder, *Associate Professor of Ecology & Evolutionary Biology*

Q&A

How did you become involved in doing research?

I applied for an REU (Research for Undergraduates) at KU last summer and worked with Dr. Mark Holder for ten weeks. It was an amazing experience!

How is the research process different from what you expected?

The amount of trial and error used in fixing everything. It's not as simple as going from point A to B. There's a lot of tear-downs in between.

What is your favorite part of doing research?

Running simulations on the supercomputer.

Performance study of supertree methods

Patrick Niehaus

INTRODUCTION

Phylogenetic trees depict the relationships among varying species [1]. A simple example of this is a family tree. When referring to Figure 1, the three most important components of a tree include: the Node(s) (denoted by squares), the Branches (denoted by lines), and the

Tips (denoted by circles). The Nodes represent the hypothetical ancestors. The Branches represent the evolution of a lineage. Lastly, the Tips represent the species being studied.

Most phylogenetic trees are estimated from DNA sequence data. A supertree is a phylogenetic tree estimate which is produced by merging smaller phylogenetic tree estimates together [2-4]. By combining several phylogenetic trees together, we can gain a better understanding of how species are related to one another. One example would be combining the human and primate phylogenetic trees.

MRP works by translating each branch of each individual tree into a matrix column; this process is repeated until a character matrix is formed [5,6]. All taxa that descend from the branch are represented in the matrix with a '1', other taxa that are found in the tree are represented with a '0', and taxa that aren't found in the tree are represented with the missing data symbol '?'. Each branch of each input tree is encoded as a series of columns. Once complete, the character matrix includes all of the phylogenetic relationships contained in the source trees. The supertree is estimated from this

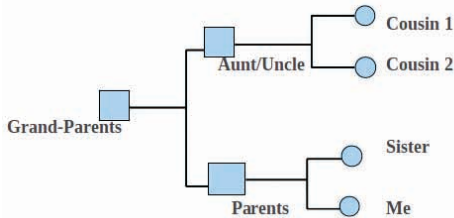


Figure 1. This figure is a visual example of a tree.

matrix using the same parsimony approach that is used when estimating a tree from two-state character data. Refer to Figure 2 for additional reference.

TAG exploits the fact that trees are actually part of a graph. Phylogenetic trees are directed and require that each node has, at most, one parent. However, TAG relaxes these requirements, and allows multiple trees to be combined into one common graph [1]. Through this process, computational time and resources are believed to be greatly minimized. Refer to Figure 3 for additional reference.

MRP is one of the more accurate supertree methods for problems with smaller taxa sets. When faced with a larger taxa set, MRP loses accuracy. TAG, however, is designed to work best with larger taxa inputs and is hypothesized to be more accurate and efficient when working with a larger taxonomy input.

METHODS

This study first requires an input taxonomy. A true tree is generated from the taxonomy by randomly resolving areas of uncertainty and then altering the tree to mimic the effects of taxonomic errors. Several input trees are generated from the true tree to mimic a set of inputs which could be obtained by several overlapping phylogenetic studies. The input trees produced sub-sample the taxa in the taxonomy, and also differ from the true tree because the simulator introduces random changes to the topology. Each input tree is then run individually through MRP and TAG as an estimate and are then compared to the true tree. Whichever estimate is closest to the true tree (e.g. the least amount of false negatives and positives) and takes the least amount of time is believed to be the more accurate and efficient algorithm. Refer to Figure 4 for additional information.

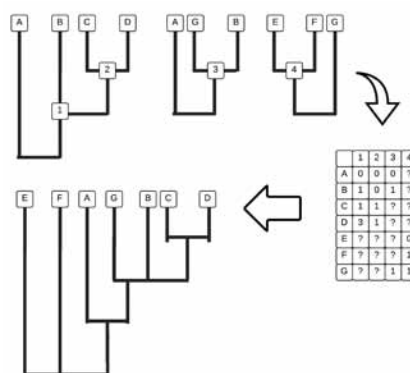


Figure 2. This figure is a visual example of the MRP process.

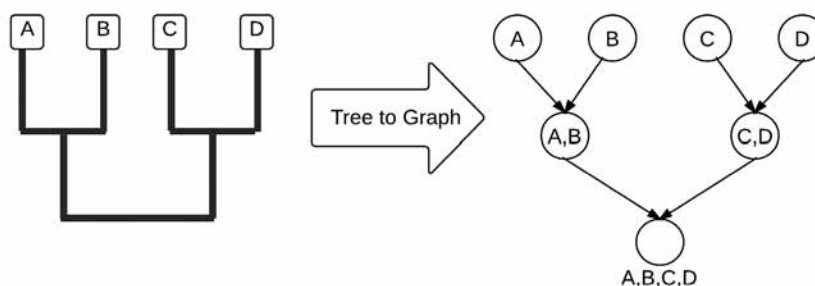


Figure 3. This figure is a visual example of the TAG process.

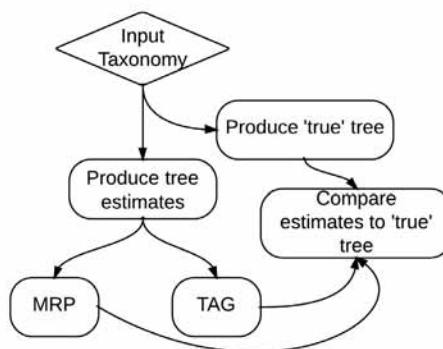


Figure 4. This figure is a visual representation of my model.

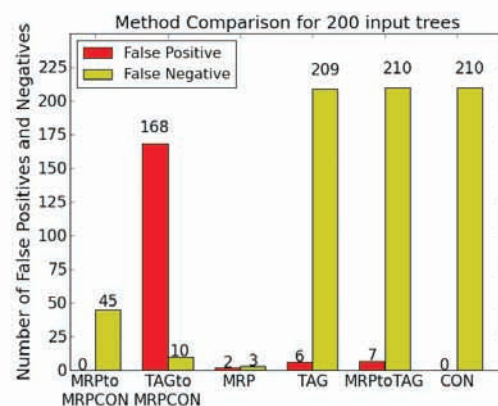
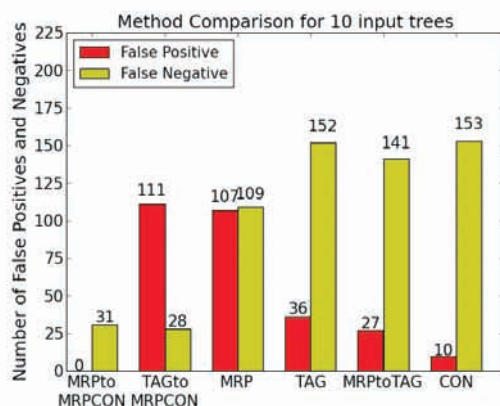
RESULTS

At this point of the study, MRP and TAG have only been compared with the primates' taxonomy, totaling around 1400 input lines. With the smaller taxonomy input, there are many differences in the outputs produced by MRP and TAG. The differences seem to be caused by false positives, false negatives, and differences between MRP and TAG to the consensus tree (CON). A false positive occurs when the input tree has a branch that isn't found in the true tree, and a false negative occurs when the true tree has a branch that isn't found in the input tree. A consensus tree is a tree that has all of the branches common to the input

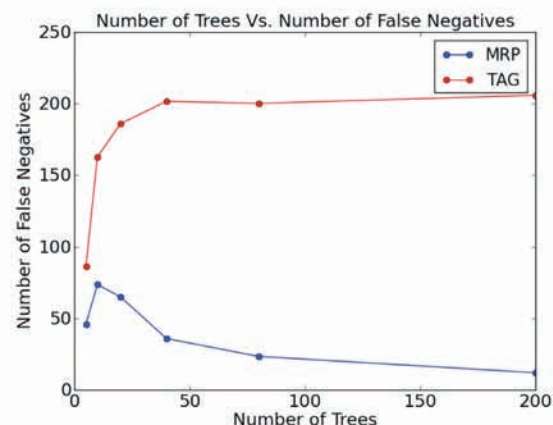
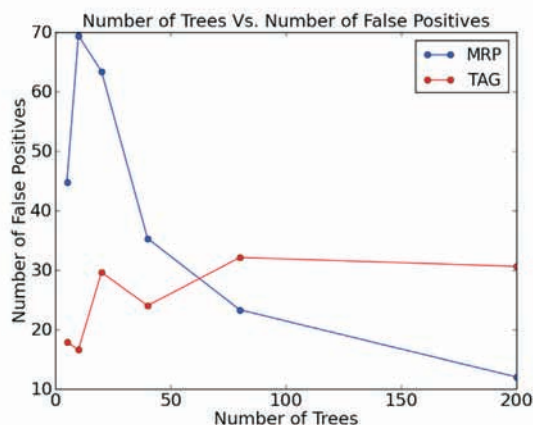
trees. This is particularly helpful in finding how similar the outputs of MRP and TAG are.

As of now, it appears that MRP far surpasses TAG for a smaller taxa input, such as the primates' taxonomy. However, Figures 5 and 6 show how the number of input trees affects the outputs of MRP and TAG in terms of false negatives and positives. MRP appears to become more effective in terms of false negatives and positives as the number of input trees increases. TAG, however, appears to become less effective as the number of trees increases, and even appears to have about the same number of false negatives/positives after around 60 input trees.

To explain why the outputs of MRP and TAG vary so greatly when the number of input trees was increased, detailed bar graphs were generated for 10 and 200 total input trees (Figures 7 and 8). When referring to Figure 7, it appears that TAG surpasses MRP in terms of false positives and false negatives. However, when looking at 'MRPtoTAG' we see that the two are returning very different outputs, and referring to 'TAGtoMRP' only confirms the varying outputs. Figure 8 shows that MRP far surpasses TAG in both false positives and false negatives, and in its comparison to the consensus tree. Therefore, it's easily apparent that MRP requires a



Figures 7 and 8. These two figures show a detailed method comparison for MRP and TAG. These two outputs were chosen at random. This was done to avoid any bias.



Figures 5 and 6. These two figures show how increasing the number of input trees affects the number of false positives and false negatives. Each dot represents an average of 10 different simulations for a given number (x) of input trees. This was done to avoid any large fluctuations in the outputs of MRP and TAG. The number of input trees (x-axis) goes in the order of 5, 10, 20, 40, 80, and 200.

larger number of input trees, but TAG, when finished, may be able to require less input trees.

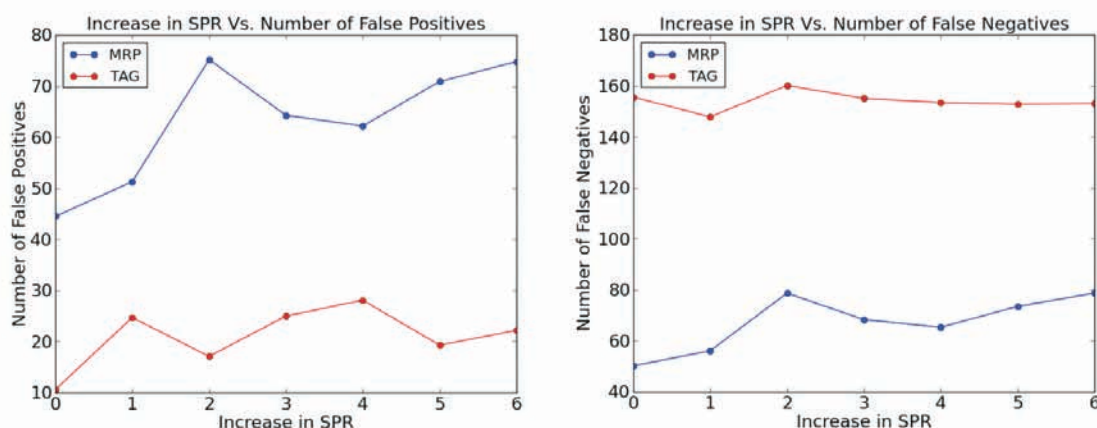
Figures 9 and 10 show what happens when we increase the SPR value for a total of 80 input trees. “SPR” stands for Sub-tree Prune Re-graph. In essence, this cuts off a branch of the tree and reconnects it elsewhere. The larger the SPR value, the further along the tree the branch is placed. Interestingly enough, this caused MRP to return more false positives as SPR was increased, but

TAG’s false positives remained below 30. In terms of false negatives, TAG remains consistent, and MRP slowly rises. At this time, we’re unsure as to why TAG returns so many false negatives despite the lower SPR value. This will be further investigated at a later stage in the study.

CONCLUSION

While MRP seems to be the more effective algorithm in this part of the study, it is important to remember that TAG was designed to work

with inputs in the form of many large overlapping phylogenetic trees with a complementary taxonomic hierarchy. It is unclear how to best simulate a realistic taxonomic input in our study system, which may result in suboptimal performance by the algorithm. In future work, MRP and TAG will be compared with a larger taxonomy (e.g. the Tree of Life, which is around 2.7 million species).



Figures 9 and 10. These two figures show how changes in SPR affect the number of false positives and false negatives. Each dot represents an increase in SPR for a total of 80 input trees over an average of 10 simulations.

This study is being conducted with the principles of open source code. All non-proprietary software is publicly viewable through the links listed below.

- <https://github.com/mtholder/supertree-study>. This link holds the code which ran the entire model illustrated in Figure 4.
- <https://github.com/OpenTreeOfLife/big-tree-collection-simulator>. This link represents the algorithm that takes the input taxonomy and creates a ‘true’ tree and the input tree(s). Refer to Figure 4 for additional information.
- <https://github.com/OpenTreeOfLife/treemachine>. This link holds the code of the TAG algorithm.

Bibliography

1. Smith SA, Brown JW, Hinchliff CE (2013) Analyzing and Synthesizing Phylogenies Using Tree Alignment Graphs. *PLoS Comput Biol* 9(9): e1003223. doi:10.1371/journal.pcbi.1003223
2. Bininda-Emonds OR, editor (2004) Phylogenetic supertrees: Combining information to reveal the Tree of Life (Computational Biology). Springer, 564 pp.
3. Davis KE, Hill J (2010) The Supertree Tool Kit. *BMC Research Notes* 3: 1–6.
4. Von Haeseler A (2012) Do we still need supertrees?. *BMC Biology* 10: 13. doi: 10.1186/preaccept-2146874722677283
5. Hill J, Davis K (2014) The Supertree Toolkit 2: a new and improved software package with a Graphical User Interface for supertree construction. *Biodiversity Data Journal* 2: e1053. doi: 10.3897/BDJ.2.e1053